Mesures de distance entre modalités de variables qualitatives; application à la classification

Hisham Abdallah, Gilbert Saporta

REVUE DE STATISTIQUE APPLIQUÉE

H. ABDALLAH
G. SAPORTA

Mesures de distance entre modalités de variables qualitatives; application à la classification

Revue de statistique appliquée, tome 51, n° 2 (2003), p. 75-90
<http://www.numdam.org/item?id=RSA_2003__51_2_75_0>

Rev. Statistique Appliquée, 2003, **LI** (2), 75-90

MESURES DE DISTANCE ENTRE MODALITÉS DE VARIABLES QUALITATIVES; APPLICATION À LA CLASSIFICATION

H. ABDALLAH[1], G. SAPORTA[2]

[1] *Université Libanaise, Faculté des sciences, Hadath. CNAM. BP 113 6175 Hamra Beyrouth.*

[2] *CNAM, Chaire de Statistique Appliquée, CEDRIC, 292 rue Saint-Martin. 75003 Paris.*

RÉSUMÉ

Regrouper des modalités appartenant aux mêmes variables ou à des variables différentes nécessite l'usage d'un indice de dissimilarité entre modalités. Nous donnons une solution à ce problème dans cet article, en nous basant sur les indices de similarité d'Ochiai, de Dice, etc. Ensuite, nous calculons des bornes pour certains indices de dissimilarité, et nous vérifions les axiomes de distance. Ceux-ci sont ensuite utilisés pour la classification ascendante hiérarchique d'un ensemble de modalités.

Mots-clés : *Variables qualitatives, modalités, distance, classification.*

ABSTRACT

Clustering categories needs a definition of dissimilarities between modalities. Our treatment will be based on Ochiai, Dice etc. indices of similarity. We give bounds and check axioms of distance for some indices of dissimilarity. We then use them to get agglomerative hierarchical clustering of a set of modalities.

Keywords : *Categorical variables, modalities, distance, classification.*

1. Dissimilarités entre modalités

1.1. Introduction

Soit un ensemble de n individus $I = \{O_1, \ldots, O_n\}$ décrits par un ensemble de p variables qualitatives $V = \{V_1, V_2, \ldots, V_p\}$ ayant chacune respectivement $m_1, m_2, \ldots, m_p$ modalités.

$\sum_{k=1}^{p} m_k = m$, où m est le nombre total des modalités.

Regrouper des modalités correspond à regrouper des propriétés. Ce regroupement nécessite la définition d'un indice de dissimilarité entre modalités j et j', noté $\delta_{jj'}$. Cet article répond à cette nécessité en introduisant la dissimilarité $\delta_{jj'} =$ Borne $-$ Intersection, où la borne est la similarité maximum entre 2 modalités j et j' et l'intersection le cardinal de $j \cap j'$.

1.2. Rappel sur les indices de similarité

Dans ce qui suit $n_{jj'}$ représente le nombre d'individus qui présentent simultanément la modalité j et la modalité j', n_j $(n_{j'})$ désigne le nombre d'individus qui possédent la modalité j (j'), $n_{j\bar{j}'}$ est le nombre d'individus possédant j et pas j', idem pour $n_{\bar{j}j'}$.

Parmi les nombreux indices de similarité qui ont été proposés [6], nous utiliserons les suivants :

Indice de Dice : $\dfrac{n_{jj'}}{\dfrac{1}{2}(n_j + n_{j'})}$

Indice d'Ochiai : $\dfrac{n_{jj'}}{\sqrt{n_j n_{j'}}}$

Indice de Kulczynski : $\dfrac{n_{jj'}}{2}\left(\dfrac{1}{n_j} + \dfrac{1}{n_{j'}}\right)$

1.3. Définitions et rappel des tableaux de base et des notations qui seront utilisées dans la suite de cet article

1.3.1. Le tableau disjonctif complet K

Le tableau K est de dimension $n \times m$, son terme général k_{ij} est défini de la façon suivante :

$$k_{ij} = \begin{cases} 1 & \text{si } i \text{ possède la modalité } j \\ 0 & \text{sinon} \end{cases}$$

Il possède les propriétés suivantes :

$$\sum_{j=1}^{m} k_{ij} = k_{i.} = p$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} k_{ij} = p \times n$$

1.3.2 Le Tableau de Burt B

Le tableau B est de dimension $m \times m$, son terme général noté $b_{jj'}$ est défini par :

$$b_{jj'} = \sum_{i=1}^{n} k_{ij} k_{ij'}.$$

On a : $b_{jj'} = n_{jj'}$, $B = K'K$.

Les termes diagonaux de la matrice de Burt sont égaux aux marges en colonnes du tableau disjonctif complet K, en effet :

$$b_{ii'} = \sum_{i=1}^{m} k_{ij} = k_{.j}.$$

1.3.3. Le tableau de Condorcet C

Le tableau C est de dimension $n \times n$, son terme général noté $c_{ii'}$ est défini par :

$$c_{ii'} = \sum_{j=1}^{m} k_{ij} k_{i'j}.$$

$c_{ii'}$ est égale au nombre de variables pour lesquelles les objets i et i' possèdent la même modalité.

On a : $C = KK'$

1.3.4. Le tableau de Burt pondéré

Le tableau de Burt pondéré, noté $\hat{B}$, est défini à partir du tableau K par pondération de la contribution par ligne d'un individu i, son terme général se présente sous la forme suivante :

$$\hat{b}_{jj'} = \sum_{i=1}^{n} \frac{k_{ij} k_{ij'}}{k_{i.}} = \frac{b_{jj'}}{p}$$

1.3.5. Le tableau de Condorcet pondéré

Le tableau de Condorcet pondéré, noté $\hat{C}$, est défini à partir du tableau K par pondération de la contribution par colonne des modalités j, son terme général s'obtient donc par la formule suivante :

$$\hat{c}_{ii'} = \sum_{j=1}^{m} \frac{k_{ij} k_{i'j}}{k_{.j}}$$

En ce qui concerne les dissimilarités on notera :

$\delta^{d}_{jj'}$ celle qui est basée sur l'indice de Dice .

$\delta^{o}_{jj'}$ celle qui est basée sur l'indice de d'Ochiai.

$\delta^{k}_{jj'}$ celle qui est basée sur l'indice de Kulczynski.

$\delta^{Union}_{jj'}$ celle qui est basée sur l'Union $(b_{jj}, b_{j'j'}) = b_{jj} + b_{j'j'} - b_{jj'}$

$\delta^{Max}_{jj'}$ celle qui est basée sur le Max$(b_{jj}, b_{j'j'})$.

$\delta^{Min}_{jj'}$ celle qui est basée sur le Min$(b_{jj}, b_{j'j'})$.

1.4. Calcul de $\delta_{jj'}$ basé sur l'indice de Dice

Sur le tableau disjonctif K, la similarité entre les individus i et i', au sens de l'indice de Dice, est :

$$S_D(i,i') = \frac{\sum_{j=1}^{m} k_{ij}k_{i'j}}{\frac{1}{2}(k_{i.} + k_{i'.})} = \frac{\sum_{j=1}^{m} k_{ij}k_{i'j}}{p} = \frac{c_{ii'}}{p}$$

Cet indice est compris entre 0 et 1. Dire que $S_D(i,i') \geqslant \frac{1}{2}$ est équivalent à dire que $c_{ii'} \geqslant \frac{p}{2}$ (il y a une majorité de variables pour lesquelles i et i' sont dans la même catégorie). Ce seuil de $\frac{1}{2}$ a été utilisé par Solomon et Fortier [7] dans leur procédure d'optimisation et correspond au coût $2c_{ii'} - p$ du critère de Condorcet. Rappelons que le critère de Condorcet (ou règle de la majorité sous contraintes) conduit à la recherche d'une partition centrale des individus qui correspond à une relation d'équivalence inconnue X telle que :

$$\max_{X} \sum_{i=1}^{n}\sum_{i'=1}^{n} (c_{ii'}x_{ii'} + \overline{c}_{ii'}\overline{x}_{ii'})$$

avec $x_{ii'} \in \{0,1\}\ \forall(i,i'), \overline{x}_{ii'} = 1 - x_{ii'}, \overline{c}_{ii'} = p - c_{ii'}$

Ce qui est équivalent, s'il n'y a pas de données manquantes, à la maximisation de la fonction de $x_{ii'}$ suivante :

$$\max_{X} \sum_{i=1}^{n}\sum_{i'=1}^{n} (2c_{ii'} - p)x_{ii'} + \sum_{i=1}^{n}\sum_{i'=1}^{n} \overline{c}_{ii'}$$

$$x_{ii'} - x_{i'i} = 0 \quad \forall(i,i') \quad \text{(symétrie)}$$

$$x_{ii'} + x_{i'i''} - x_{ii''} \leqslant 1 \quad \forall(i,i',i'') \quad \text{(transitivité)}$$

De façon équivalente, on peut calculer à partir du tableau K la similarité selon l'indice de Dice entre les profils des modalités j et j' :

$$S_D(j,j') = \frac{\sum_{i=1}^{n} k_{ij}k_{ij'}}{\frac{1}{2}(k_{.j} + k_{.j'})} = \frac{b_{jj'}}{\frac{1}{2}(b_{jj} + b_{j'j'})}, \text{ indice qui varie entre 0 et 1.}$$

Dire que l'on est supérieur à la borne indicielle se traduit par :

$$S_D(j,j') \geqslant \frac{1}{2} \Longrightarrow b_{jj'} \geqslant \frac{b_{jj'} + b_{j'j'}}{4} \Longrightarrow 2b_{jj'} \geqslant \frac{b_{jj} + b_{j'j'}}{2}$$

À la relation d'équivalence inconnue $X = \{x_{ii'}\}$ associée au critère de Condorcet, on fait correspondre une relation d'équivalence inconnue $Y = \{y_{jj'}\}$ associée au critère appelé critère de Burt .

Au coût de la fonction économique du critère de Condorcet $(2c_{ii'} - p)$ ou plus généralement $(c_{ii'} - \overline{c}_{ii'})$ on associera le coût $\left(2b_{jj'} - \frac{1}{2}(b_{jj} + b_{j'j'})\right)$ et plus généralement $(b_{jj'} - \delta^d_{jj'})$, pour la fonction économique du critère de Burt.

Autrement dit : $b_{jj'} - \delta^d_{jj'} + 2b_{jj'} - \frac{1}{2}(b_{jj} + b_{j'j'})$

On en déduit donc la valeur de $\delta^d_{jj'}$:

$$\delta^d_{jj'} = \frac{b_{jj} + b_{j'j'}}{2} - b_{jj'} = \frac{1}{2}\sum_{i=1}^{n}(k_{ij}\overline{k}_{ij'} + k_{ij'}\overline{k}_{ij}) = \frac{1}{2}\sum_{i=1}^{n}(k_{ij} - k_{ij'})^2$$

avec $\overline{k}_{ij} = 1 - k_{ij}$

Nous remarquons ici que la similarité maximum entre les modalités j et j' est la moyenne arithmétique de $(b_{jj}, b_{j'j'})$.

À partir de cette expression de $\delta^d_{jj'}$, nous pouvons construire la formulation mathématique du problème d'optimisation associé au «critère de Burt», c'est-à-dire le moyen d'obtenir une partition Y des modalités (sans fixation du nombre de classes) solution du programme linéaire à variables bivalentes suivant :

$$\max_{Y} B(Y)$$

avec $B(Y) = \sum_{j=1}^{m}\sum_{j'=1}^{m}(b_{jj'}y_{jj'} + \delta^d_{jj'}\overline{y}_{jj'})$

Ce qui est équivalent, à la maximisation de la fonction de $y_{jj'}$ suivante :

$$\max_{y} \sum_{j=1}^{m}\sum_{j'=1}^{m}\left(2b_{jj'} - \frac{1}{2}(b_{jj} + b_{j'j'})\right)y_{jj'} + \sum_{j=1}^{m}\sum_{j'=1}^{m}\delta^d_{jj'}$$

$$y_{jj'} - y_{j'j} = 0 \quad \forall(j,j') \quad \text{(symétrie)}$$

$$y_{jj'} + y_{j'j''} - y_{jj''} \leqslant 1 \quad \forall(j,j',j'') \quad \text{(transitivité)}$$

$$y_{jj'} \in \{0,1\} \quad \forall(j,j') \quad \text{(binarité)}$$

De façon analogue, on trouve les autres valeurs de $\delta_{jj'}$:

$$\delta^0_{jj'} = \sqrt{\sum_{i=1}^{n} k_{ij}} \sqrt{\sum_{i=1}^{n} k_{ij'}} - \sum_{i=1}^{n} k_{ij} k_{ij'}$$

$$\delta^k_{jj'} = \frac{2 b_{jj} b_{j'j'}}{b_{jj} + b_{j'j'}} - b_{jj'}$$

$$\delta^{Max}_{jj'} = Max(b_{jj}, b_{j'j'}) - b_{jj'} \tag{1.4.1}$$

$$\delta^{Min}_{jj'} = Min(b_{jj}, b_{j'j'}) - b_{jj'} \tag{1.4.2}$$

$$\delta^{union}_{jj'} = Union(b_{jj}, b_{j'j'}) - b_{jj'} \tag{1.4.3}$$

Concernant ces indices de dissimilarité, les similarités maximales entre les modalités j et j' sont respectivement : la moyenne géométrique de $(b_{jj}, b_{j'j'})$, la moyenne harmonique de $(b_{jj}, b_{j'j'})$, le maximum de $(b_{jj}, b_{j'j'})$, le minimum de $(b_{jj}, b_{j'j'})$, l'union de $(b_{jj}, b_{j'j'})$.

1.5. Relation entre $\delta^o_{jj'}$ et $\delta^d_{jj'}$

$$\delta^o_{jj'} = \delta^d_{jj'} - \frac{1}{2}(\sqrt{k_{.j}} - \sqrt{k_{.j'}})^2$$

Preuve.

En effet :

$$\left(\sqrt{\sum_{i=1}^{n} k_{ij}} - \sqrt{\sum_{i=1}^{n} k_{ij'}} \right)^2 = \sum_{i=1}^{n} k_{ij} + \sum_{i=1}^{n} k_{ij'} - 2\sqrt{\sum_{i=1}^{n} k_{ij} \sum_{i=1}^{n} k_{ij'}} \tag{1.5.1}$$

$$\Longrightarrow \sqrt{\sum_{i=1}^{n} k_{ij} \sum_{i=1}^{n} k_{ij'}} = \frac{1}{2}\left[\sum_{i=1}^{n} k_{ij} + \sum_{i=1}^{n} k_{ij'} \right] - \frac{1}{2}\left(\sqrt{\sum_{i=1}^{n} k_{ij}} - \sqrt{\sum_{i=1}^{n} k_{ij'}} \right)^2$$

$\delta^o_{jj'}$ devient alors :

$$\delta^o_{jj'} = \frac{1}{2}\left[\sum_{i=1}^{n} k_{ij} + \sum_{i=1}^{n} k_{ij'} - 2\sum_{i=1}^{n} k_{ij} k_{ij'} \right] - \frac{1}{2}\left(\sqrt{\sum_{i=1}^{n} k_{ij}} - \sqrt{\sum_{i=1}^{n} k_{ij'}} \right)^2$$

$$= \frac{1}{2}\left(\sum_{i=1}^{n} (k_{ij} - k_{ij'}) \right)^2 - \frac{1}{2}\left(\sqrt{\sum_{i=1}^{n} k_{ij}} - \sqrt{\sum_{i=1}^{n} k_{ij'}} \right)^2$$

$$= \frac{1}{2}\left[\sum_{i=1}^{n} \left[k_{ij} \overline{k}_{ij'} + k_{ij'} \overline{k}_{ij} \right] \right] - \frac{1}{2}(\sqrt{k_{.j}} - \sqrt{k_{.j'}})^2$$

1.6. Relation entre $\delta^k_{jj'}$ et $\delta^d_{jj'}$

$$\delta^k_{jj'} = \delta^d_{jj'} - \frac{1}{2}\frac{(k_{.j} - k_{.j'})^2}{k_{.j} + k_{.j'}}$$

Preuve.

$$\left(\frac{k_{.j}}{k_{.j} + k_{.j'}} + \frac{k_{.j'}}{k_{.j} + k_{.j'}}\right)^2 = \frac{k^2_{.j}}{(k_{.j} + k_{.j'})^2} + \frac{k^2_{.j'}}{(k_{.j} + k_{.j'})^2} + 2\,\frac{k_{.j}k_{.j'}}{(k_{.j} + k_{.j'})^2}$$

Sachant que :

$$\frac{k_{.j}}{k_{.j} + k_{.j'}} + \frac{k_{.j'}}{k_{.j} + k_{.j'}} = 1$$

On obtient :

$$2\,\frac{k_{.j}k_{.j'}}{(k_{.j} + k_{.j'})^2} = 1 - \left(\frac{k^2_{.j}}{(k_{.j} + k_{.j'})^2} + \frac{k^2_{.j'}}{(k_{.j} + k_{.j'})^2}\right)$$

Donc :

$$\begin{aligned}
\delta^k_{jj'} &= k_{.j} + k_{.j'} - \left(\frac{k^2_{.j}}{(k_{.j} + k_{.j'})^2} + \frac{k^2_{.j'}}{(k_{.j} + k_{.j'})^2}\right)(k_{.j} + k_{.j'}) - \sum_{i=1}^{n} k_{ij}k_{ij'} \\
&= \sum_{i=1}^{n} k_{ij} + \sum_{i=1}^{n} k_{ij'} - \frac{1}{k_{.j} + k_{.j'}}(k^2_{.j} + k^2_{.j'}) - \sum_{i=1}^{n} k_{ij}k_{ij'} \\
&= \frac{1}{2}\left(\sum_{i=1}^{n}(k_{ij} + k_{ij'} - 2k_{ij}k_{ij'})\right) + \frac{1}{2}\sum_{i=1}^{n}(k_{ij} + k_{ij'}) - \frac{1}{k_{.j} + k_{.j'}}(k^2_{.j} + k^2_{.j'}) \\
&= \frac{1}{2}\sum_{i=1}^{n}\left[k_{ij}\bar{k}_{ij'} + k_{ij'}\bar{k}_{ij}\right] - \frac{1}{2}\,\frac{(k_{.j} - k_{.j'})^2}{k_{.j} + k_{.j'}}
\end{aligned}$$

Pour les autres indices de dissimilarité, on trouve les relations suivantes :

$\delta^{Max}_{jj'} = \delta^d_{jj'} + \frac{1}{2}|k_{.j} - k_{.j'}|$.

Afin d'obtenir cette relation, il suffit de remplacer le maximum de $(b_{jj}, b_{j'j'})$ par sa valeur dans la relation (1.4.1).

$\delta^{Min}_{jj'} = \delta^d_{jj'} - \frac{1}{2}|k_{.j} - k_{.j'}|$.

De même, pour démontrer cette relation, on remplace le minimum de $(b_{jj}, b_{j'j'})$ par sa valeur dans la relation (1.4.2).

$\delta^U_{jj'} = 2\delta^d_{jj'}$

En effet, pour vérifier cette égalité, on remplace l'Union de $(b_{jj}, b_{jj'}) = b_{jj} + b_{jj'} - b_{jj'}$ par sa valeur dans la relation (1.4.3)

2. Majoration de quelques indices

2.1 Majoration de $\delta^0_{jj'}$

Soit $\delta^o_{jj'} = \sqrt{\sum_{i=1}^{n} k_{ij}}\sqrt{\sum_{i=1}^{n} k_{ij'}} - \sum_{i=1}^{n} k_{ij}k_{ij'}$

Posons $I = \{i; k_{ij} = 1\}$ et $I' = \{i; k_{ij'} = 1\}$, alors on obtient :

$$\delta^o_{jj'} = \sqrt{|I|}\sqrt{|I'|} - |I \cap I'| \quad \text{avec} \quad |I \cap I'| = n_{jj'}$$

calculons $\sup_{I,I' \subset \{1,\ldots,n\}} \left[\sqrt{|I|}\sqrt{|I'|} - |I \cap I'|\right] = \sup\ \delta^o_{jj'}$

1er cas : $I \cap I' = \varnothing$ alors on obtient :

$\sup_{I \cap I' = \varnothing} \delta^o_{jj'} \leqslant \sup_{\substack{0 \prec s,y \\ x+y \leqslant n}} \sqrt{xy} = \sup_{\substack{0 \prec x,y \\ x+y=n}} \sqrt{xy} = \alpha_n$ où $\alpha_n = \sqrt{\frac{n}{2}\frac{n}{2}} = \frac{n}{2}$, cette borne peut être atteinte si n est pair. Si n est impair, on a une majoration plus précise égal à $\frac{\sqrt{n^2-1}}{2}$ atteinte pour $x = \frac{(n-1)}{2}$, $y = \frac{(n+1)}{2}$ ou l'inverse.

2e cas : $|I \cap I'| = n_{jj'}$ où $0 \leqslant n_{jj'} \leqslant n$ soit $L = I \cap I'$
et $n_{jj'} = |L|$ avec $I = I_1 \cup L$ et $I' = I'_1 \cup L$
où I_1, I'_1 sont deux parties disjointes d'un ensemble ayant $n - n_{jj'}$ éléments
donc :
$\delta^o_{jj'} = \sqrt{n_{jj'} + |I_1|}\sqrt{n_{jj'} + |I'_1|} - n_{jj'}$ et,

$$\sup_{\substack{I,I' \subset \{1,\ldots,n\} \\ |I \cap I'| = n_{jj'}}} \delta^o_{jj'} \leqslant \sup_{\substack{0 \leqslant x \\ 0 \leqslant y \\ x+y=n-n_{jj'}}} \sqrt{n_{jj'}+x}\sqrt{n_{jj'}+y} - n_{jj'}$$

$$= \sup_{\substack{0 \leqslant x \\ 0 \leqslant y \\ x+y=n-n_{jj'}}} \sqrt{n_{jj'}+x}\sqrt{n_{jj'}+y} - n_{jj'} = \beta_n$$

$\left(x = y = \dfrac{n - n_{jj'}}{2} \Longrightarrow x + n_{jj'} = \dfrac{n + n_{jj'}}{2}\right)$

donc $\beta_n = \dfrac{n + n_{jj'}}{2} - n_{jj'} = \dfrac{n - n_{jj'}}{2}$

d'où $\boxed{0 \leqslant \delta^o_{jj'} \leqslant \dfrac{n - n_{jj'}}{2}}$ borne atteinte si $n - n_{jj'}$ est pair.

Sinon ($n - n_{jj'}$ est impair), on a une borne plus précise, à savoir $\dfrac{\sqrt{(n + n_{jj'})^2 - 1}}{2} - n_{jj'}$ atteinte pour $x = \dfrac{(n - n_{jj'} - 1)}{2}$, $y = \dfrac{(n - n_{jj'} + 1)}{2}$ ou l'inverse.

2.2. Majoration de la somme des $\hat{\delta}^o_{jj'}$

Nous nous situons dans l'espace des modalités J. Considérons le nuage des profils colonnes du tableau $K\ N(J) = \{(V_j, \mu_j)/j \in J\}$ dans un espace muni de la métrique du «khi-deux» où μ_j représente la masse affectée à chaque modalité : $\mu_j = f_{.j} = \dfrac{k_{.j}}{np}$.

L'inertie de ce nuage vaut ζ [3] :

$$\zeta = \frac{1}{2np^2} \sum_{j=1}^{m} \sum_{j'=1}^{m} [k_{.j} + k_{.j'} - 2b_{jj'}] = \frac{1}{np} \sum_{j=1}^{m} \sum_{j'=1}^{m} \left(\frac{k_{.j} + k_{.j'}}{2p} - \hat{b}_{jj'} \right)$$
$$= \frac{m}{p} - 1$$

or $\hat{\delta}^o_{jj'} = \sqrt{\hat{b}_{jj} \hat{b}_{j'j'}} - \hat{b}_{jj'} = \dfrac{1}{p} \sqrt{\sum_{i=1}^{n} k_{ij} \sum_{i=1}^{n} k_{ij'}} - \hat{b}_{jj'}$

En utilisant la relation (1.5.1) on obtient :

$$\hat{\delta}^o_{jj'} = \left(\frac{k_{.j} + k_{.j'}}{2p} - \hat{b}_{jj'} \right) - \frac{1}{2p} (\sqrt{k_{.j}} - \sqrt{k_{.j'}})^2$$
$$\Longrightarrow \frac{k_{.j} + k_{.j'}}{2p} - \hat{b}_{jj'} = \hat{\delta}^o_{jj'} + \frac{1}{2p} (\sqrt{k_{.j}} - \sqrt{k_{.j'}})^2$$
$$\Longrightarrow \zeta = \frac{1}{np} \sum_{j} \sum_{j'} \left(\hat{\delta}^o_{jj'} + \frac{1}{2p} (\sqrt{k_{.j}} - \sqrt{k_{.j'}})^2 \right) = \frac{m}{p} - 1$$

d'où :

$$\sum_j \sum_{j'} \hat{\delta}^o_{jj'} = np\left(\frac{m}{p} - 1\right) - \frac{1}{2p}\sum_j \sum_{j'} \left((\sqrt{k_{.j}} - \sqrt{k_{.j'}})^2\right)$$

$$\Longrightarrow \sum_j \sum_{j'} \hat{\delta}^o_{jj'} = nm - np - \frac{1}{2p}(npm + npm) + \frac{1}{p}\sum_j \sum_{j'} \sqrt{k_{.j}}\sqrt{k_{.j'}}$$

$$= \frac{1}{p}\sum_j \sum_{j'} \sqrt{k_{.j}}\sqrt{k_{.j'}} - np$$

$$= \frac{1}{p}\left(\sum_{j=1}^{m} \sqrt{k_{.j}}\right)^2 - np \leqslant \frac{1}{p}\left(m\left(\sum_{j=1}^{m} k_{.j}\right)\right) - np \text{ (Inégalité de Holder)}$$

$$= \frac{1}{p}(pnm) - np = nm - np$$

Donc

$$\boxed{\hat{\delta}^o_{..} = \sum_j \sum_{j'} \hat{\delta}^o_{jj'} \leqslant nm - np}$$

3. Propriétés métriques des indices $\delta_{jj'}$ précédemment définis

Les résultats concernant la stucture géométrique des indices $\delta_{jj'}$ sont liés aux répartitions de $k_{.j}$. Tous ces indices sont des indices de dissimilarité puisque $\delta_{jj} = 0$ et $\delta_{jj'} = \delta_{j'j}$. Nous allons montrer que certains d'entre eux sont des distances.

1er cas : Si $k_{.j} = k_{.j'}$, tous les $\delta_{jj'}$ sont équivalents et correspondent à des distances euclidiennes.

2e cas : Si $k_{.j} \neq k_{.j'}$ nous allons voir que $\delta^d_{jj'}$, $\delta^{Union}_{jj'}$ et $\delta^{Max}_{jj'}$ sont des distances.

En effet $\delta^d_{jj'} = \frac{1}{2}\left(\sum_{i=1}^n (k_{ij} - k_{ij'})^2\right)$ est une distance euclidienne, donc $\delta^{Union}_{jj'} = 2\delta^d_{jj'}$ est aussi une distance euclidienne.

On va montrer maintenant que $\delta^{Max}_{jj'}$ vérifie l'inégalité triangulaire, et est donc une distance.

Pour $\delta^{Max}_{jj'} = \delta^d_{jj'} + \frac{1}{2}|b_{jj} - b_{j'j'}|$, vérifions que $\delta^{Max}_{jj'} - \delta^{Max}_{jj''} - \delta^{Max}_{j''j'} \leqslant 0$:

$$\delta^{Max}_{jj'} - \delta^{Max}_{jj''} - \delta^{Max}_{j''j'}$$
$$= \delta^d_{jj'} - \delta^d_{jj''} - \delta^d_{j''j'} + \frac{1}{2}\Big[|b_{jj} - b_{j'j'}| - |b_{jj} - b_{j''j''}| - |b_{j''j''} - b_{j'j'}|\Big].$$

Or $\delta^d_{jj'} - \delta^d_{jj'} - \delta^d_{j''j'} \leqslant 0$ puisque $\delta^d_{jj'}$ est une distance. Par ailleurs, on sait que $|b_{jj} - b_{j'j'}| \leqslant |b_{jj} - b_{j''j''}| + |b_{j''j''} - b_{j'j'}|$. D'où $\delta^{Max}_{jj'} - \delta^{Max}_{jj''} - \delta^{Max}_{j''j'} \leqslant 0$ donc $\delta^{Max}_{jj'}$ est une distance. Comme $\delta^{Max}_{jj'}$ n'est pas engendrée par un produit scalaire, $\delta^{Max}_{jj'}$ est une distance non euclidienne.

Pour les indices d'Ochiai $\delta^o_{jj'}$, Kulczynski $\delta^k_{jj'}$ et $\delta^{Min}_{jj'}$ on peut trouver des cas où l'inégalité triangulaire n'est pas vérifiée : ce sont donc simplement des dissimilarités.

3.1. Synthèse

Le tableau ci-après présente, sous forme synthétique, les cas correspondant aux différentes valeurs de $\delta_{jj'}$.

Notation	**Similarité**	**valeurs de** δ_{jj}	**Propriété**
$\delta^d_{jj'}$	Indice de Dice : moyenne arithmétique	$\frac{1}{2}\sum_{i=1}^{n}(k_{ij}\overline{k}_{ij'} + k_{ij}\overline{k}_{ij})$	est une distance euclidienne
$\delta^o_{jj'}$	Indice d'Ochiai : moyenne géométrique	$\delta^d_{jj'} - \frac{1}{2}(\sqrt{k_{.j}} - \sqrt{k_{.j'}})^2$	est un indice de dissimilarité
$\delta^k_{jj'}$	Indice de Kulczynski : moyenne harmonique	$\delta^d_{jj'} - \frac{1}{2}\frac{(k_{.j} - k_{.j'})^2}{(k_{.j} + k_{.j'})}$	est un indice de dissimilarité
$\delta^{Max}_{jj'}$	$Max(k_{.j}, k_{.j'})$	$\delta^d_{jj'} + \frac{1}{2}\lvert k_{.j} - k_{.j'}\rvert$	est une distance non euclidienne
$\delta^{Min}_{jj'}$	$Min(k_{.j}, k_{.j'})$	$\delta^d_{jj'} - \frac{1}{2}\lvert k_{.j} - k_{.j'}\rvert$	est un indice de dissimilarité
$\delta^{Union}_{jj'}$	$Union(k_{.j}, k_{.j'})$	$2\delta^d_{jj'}$ euclidienne	est une distance

4. Exemple illustratif

Reprenons le classique exemple des «canidés» [6].

Les données du tableau (4.1) décrivent les caractéristiques de 27 races de chiens au moyen de variables qualitatives, les 6 premières considérées comme actives la septième «fonction» comme supplémentaire : ses trois modalités sont «compagnie», «chasse», «utilité».

TABLEAU 4.1

	Taille	Poids	Veloc.	Intel.	Affect.	Agre.	Fonction
1. Beauceron	3	2	3	3	2	2	3
2. Basset	1	1	1	1	1	2	2
3. B.Allemand	3	2	3	3	2	2	3
4. Boxer	2	2	2	2	2	2	1
5. Bull-Dog	1	1	1	2	2	1	1
6. Bull-Mastiff	3	3	1	3	1	2	3
7. Caniche	1	1	2	3	2	1	1
8. Chihuahua	1	1	1	1	2	1	1
9. Cocker	2	1	1	2	2	2	1
10. Colley	3	2	3	2	2	1	1
11. Dalmatien	2	2	2	2	2	1	1
12. Doberman	3	2	3	3	1	2	3
13. D.Allemand	3	3	3	1	1	2	3
14. E.Breton	2	2	2	3	2	1	2
15. E.français	3	2	2	2	1	1	2
16. Fox-Hound	2	2	3	1	1	2	2
17. Fox-Terrier	1	1	2	2	2	2	1
18. G.B.de Gasco.	3	2	2	1	1	2	2
19. Labrador	2	2	2	2	2	1	2
20. Levrier	3	2	3	1	1	1	2
21. Mastiff	3	3	1	1	1	2	3
22. Pekinois	1	1	1	1	2	1	1
23. Pointer	3	2	3	3	1	1	2
24. Saint-Bernard	3	3	1	2	1	2	3
25. Setter	3	2	3	2	1	1	2
26. Teckel	1	1	1	2	2	1	1
27. Terre-Neuve	3	3	1	2	1	1	3

où :

La Taille a 3 modalités : 1 : Ta1-petite, 2 : Ta2-moyenne,3 : Ta3-grande.

Le Poids a 3 modalités : 1 : Po1-petit, 2 : Po2-moyen, 3 : Po3-lourd.

La vélocité a 3 modalités : 1 : Ve1-lent, 2 : Ve2-assez rapide, 3 : Ve3-très rapide.

L'intelligence a 3modalités : 1 : In1-médiocre, 2 : In2-moyenne, 3 : In3-forte.

L'affection a 2 modalités : 1 : Af1-peu affectueux, 2 : Af2-affectueux.

L'agressivité a 2 modalités : 1 : Ag1-peu agressif, 2 : Ag2-agressif.

L'A.F.C.M.du tableau 4.1 donne la représentation grahique suivante (*cf.* Figure 4.1) :

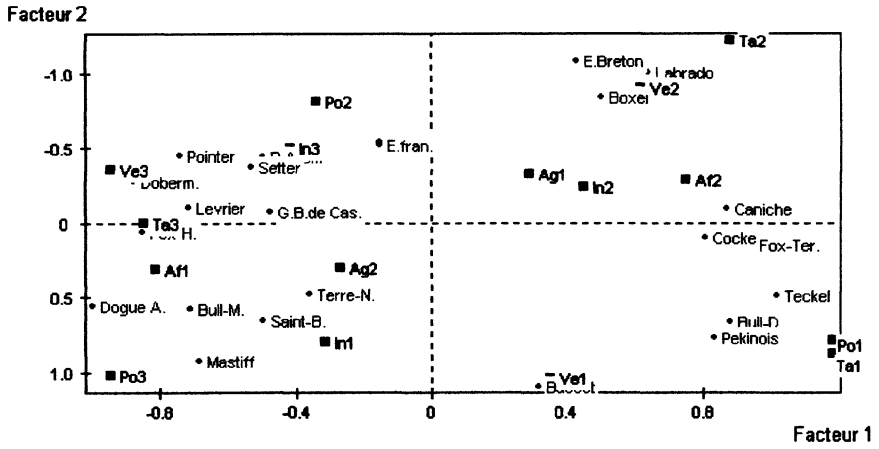

FIGURE 4.1

4.1. Classification hiérarchique

À partir du tableau disjonctif déduit du tableau 4.1, on va réaliser une classification ascendante hiérarchique entre les modalités, en utilisant $\delta^d_{jj'}$, comme un indice de distance entre ces modalités.

Le tableau de distances associé est donné dans le tableau 4.2.

Ensuite, à partir de ce tableau, on aboutit à l'arbre de classification suivante (méthode de Ward) (*cf.* Figure 4.2).

L'arbre de classification met en évidence 5 classes obtenues en coupant l'arbre à l'indice de niveau 3.2 :

La classe 1 regroupe 3 modalités : {Ta1, Po1, Ve1}.

La classe 2 regroupe 3 modalités : {Po3, Af1, In1}.

La classe 3 regroupe 2 modalités : {Ta2, Ve2}.

La classe 4 regroupe 3 modalités : {In2, Ag1, Af2}.

La classe 5 regroupe 5 modalités : {Ta3, Ve3, Po2, In3, Ag2}.

On constate un bon accord entre la classification hiérarchique et l'A.F.C.M.

TABLEAU 4.2
Tableau de distances entre modalités

	Ta1	Ta2	Ta3	Po1	Po2	Po3	Ve1	Ve2	Ve3	In1	In2	In3	Af1	Af2	Ag1	Ag2
Ta1	0	6	11	0.5	10.5	6	2.5	7.5	8	5	6.5	6	9	4.5	5.5	8
Ta2	6	0	10	5.5	5.5	5	6.5	2.5	7	6.5	4.5	5	9	4.5	6.5	7
Ta3	11	10	0	11.5	4.5	5	8.5	9.5	3	6.5	8.5	6	3	10.5	8.5	5
Po1	0.5	5.5	11.5	0	11	6.5	3	6	8.5	5	6	6.5	10.5	4	6	7.5
Po2	10.5	5.5	4.5	11	0	9.5	12	5	3.5	5	7	8.5	6.5	7	6	7.5
Po3	6	5	5	6.5	9.5	0	3.5	6.5	6	4.5	6.5	5	4	9.5	8.5	5
Ve1	2.5	6.5	8.5	3	12	3.5	0	9	9.5	5	6	7.5	6.5	7	7	6.5
Ve2	7.5	2.5		9.56	5	6.5	9	0	8.5	7	5	5.5	8.5	5	6	7.5
Ve3	8	7	3	8.5	3.5	6	9.5	8.5	0	5.5	8.5	4	5	8.5	7.5	6
In1	5	6.5	6.5	5	5	4.5	5	7	5.5	0	10	7.5	4.5	9	8	5.5
In2	6.5	4.5	8.5	6	7	6.5	6	5	8.5	10	0	9.5	8.5	5	5	8.5
In3	6	5	6	6.5	8.5	5	7.5	5.5	4	7.5	9.5	0	7	6.5	7.5	6
Af1	9	9	3	10.5	6.5	4	6.5	8.5	5	4.5	8.5	7	0	13.5	8.5	5
Af2	4.5	4.5	10.5	4	7	9.5	7	5	8.5	9	5	6.5	13.5	0	5	8.5
Ag1	5.5	6.5	8.5	6	6	8.5	7	6	7.5	8	5	7.5	8.5	5	0	13.5
Ag2	8	7	5	7.5	7.5	5	6.5	7.5	6	5.5	8.5	6	5	8.5	13.5	0

4.2. Recherche d'une partition à partir du critère de Burt

En appliquant le critère de Burt sur l'exemple des « canidés » déjà cité ci-dessus, on obtient la partition en 7 classes suivante :

Classe 1 : Ta1, Po1, Ve1.

Classe 2 : Ta2, Ve2.

Classe 3 : Ta3, Po2, Ve3, Af1, Ag2.

Classe 4 : Po3.

Classe 5 : In1.

Classe 6 : In2, Ag1,Af2

Classe 7 : In3.

En comparant cette partition à celle effectuée en classification ascendante hiérarchique, on remarque qu'il y a 3 classes invariantes, d'autres subissent un changement.

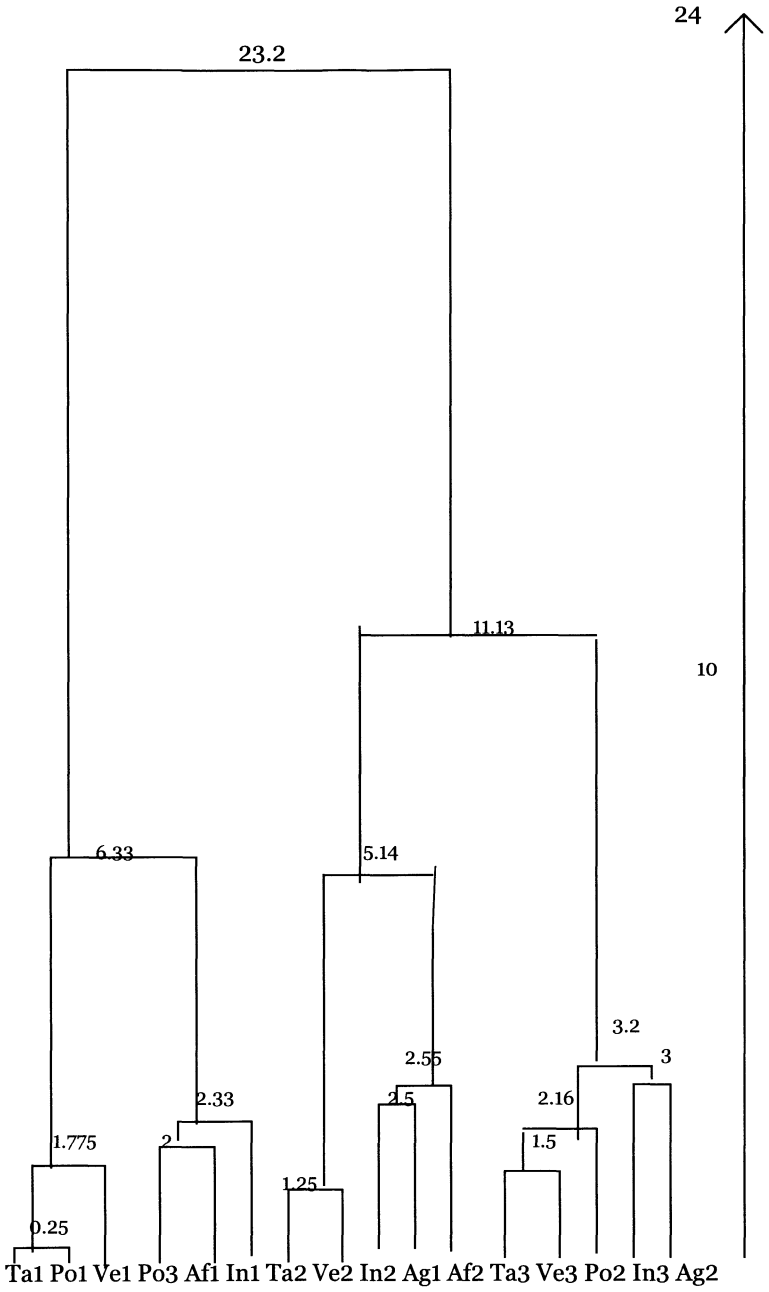

FIGURE 4.2
Classification Ascendante Hiérarchique

5. Conclusion

Nous avons vu que pour définir les désaccords $\delta_{jj'}$ entre les modalités j et j', il a fallu introduire la notion de borne où la borne est définie comme la similarité maximum entre les deux profils j et j'.

Il y a plusieurs possibilités pour choisir cette borne, nous avons étudié quelques unes parmi les plus utilisées, et selon la borne que nous avons choisie, nous avons obtenu différentes valeurs de $\delta_{jj'}$, celles-ci peuvent s'exprimer en fonction de $\delta^d_{jj'}$, $k_{.j}$ et $k_{.j'}$ autrement dit :

$\delta^*_{jj'} = \delta^d_{jj'} + f(k_{.j}, k_{.j'})$

$\delta^*_{jj'}$ désignant : $\delta^{Max}_{jj'}$, $\delta^{Min}_{jj'}$, $\delta^o_{jj'}$, $\delta^k_{jj'}$

$2f(k_{.j}, k_{.j'})$ étant au signe prés, et suivant le cas égal à $|k_{.j} - k_{.j'}|$ ou $(\sqrt{k_{.j}} - \sqrt{.j'})^2$ ou $\dfrac{(k_{.j} - k_{.j'})^2}{k_{.j} + k_{.j'}}$.

D'autre part, nous avons vu que nous pouvions partager les $\delta_{jj'}$ en deux catégories différentes : une catégorie où les $\delta_{jj'}$ sont de simples indices de dissimilarité; une autre catégorie où les $\delta_{jj'}$ sont des distances. En particulier, celles qui sont euclidiennes sont les plus utiles en pratique car elles permettent d'utiliser la méthode de Ward en classification ascendante hiérarchique, c'est le cas de la distance basée sur l'indice de Dice.

Références

[1] Abdallah H., (1996), Application de l'analyse relationnelle pour classifier descripteurs et modalités en mode discrimination,Thèse de l'Université Pierre et Marie Curie , Paris 6.

[2] L. Lebart, A. Morineau, M. Piron, (1995), Statistique exploratoire multidimensionnelle, Dunod, Paris.

[3] Marcotorchino F., (1991), L'analyse factorielle-relationnelle (parties 1 et 2). Etude du centre scientifique IBM France.

[4] Marcotorchino F., Michaud P., (1979), Optimisation en analyse ordinale des données, Masson, Paris.

[5] Meulman J., (1986), A distance approach to nonlinear multivariate analysis, DSWO Press, Leiden.

[6] Saporta G., (1990), Probabilités, Analyse des données et Statistique, éditions Technip.

[7] Solomon H. et Fortier J., (1966), «Clustering Procedures», Multivariate Analysis, P. Krishnaiah Editor, Academic Press, New york.